PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models

Year: 2024 | *Citations: 188* | *Authors: Fanxu Meng, Zhaohui Wang, Muhan Zhang*

Abstract

To parameter-efficiently fine-tune (PEFT) large language models (LLMs), the low-rank adaptation (LoRA) method approximates the model changes $\Delta W \in \mathbb{R}^{m \times n}$ through the product of two matrices $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{r \times n}$, where $r \ll \min(m, n)$, A is initialized with Gaussian noise, and B with zeros. LoRA freezes the original model W and updates the"Noise&Zero"adapter, which may lead to slow convergence. To overcome this limitation, we introduce Principal Singular values and Singular vectors Adaptation (PiSSA). PiSSA shares the same architecture as LoRA, but initializes the adaptor matrices A and B with the principal components of the original matrix W, and put the remaining components into a residual matrix $W^{res} \in \mathbb{R}^{m \times n}$ which is frozen during fine-tuning. Compared to LoRA, PiSSA updates the principal components while freezing the"residual"parts, allowing faster convergence and enhanced performance. Comparative experiments of PiSSA and LoRA across 12 different models, ranging from 184M to 70B, encompassing 5 NLG and 8 NLU tasks, reveal that PiSSA consistently outperforms LoRA under identical experimental setups. On the GSM8K benchmark, Mistral-7B fine-tuned with PiSSA achieves an accuracy of 72.86%, surpassing LoRA's 67.7% by 5.16%. Due to the same architecture, PiSSA is also compatible with quantization to further reduce the memory requirement of fine-tuning. Compared to QLoRA, QPiSSA exhibits smaller quantization errors in the initial stages. Fine-tuning LLaMA-3-70B on GSM8K, QPiSSA attains an accuracy of 86.05%, exceeding the performances of QLoRA at 81.73%. Leveraging a fast SVD technique, PiSSA can be initialized in only a few seconds, presenting a negligible cost for transitioning from LoRA to PiSSA. Code is available at https://github.com/GraphPKU/PiSSA.